WHAT IS CLAIMED IS:

1. A computer-implemented method of analyzing a sequence of amino acids, comprising;

(a) designating each amino acid within the sequence with a symbol, wherein an amino acid is designated a first symbol if it is a member of a predetermined set of amino acids, and a second symbol different from the first symbol if the amino acid is not a member of the predetermined set, thereby producing a sequence of symbols;

(b) determining which signals of the symbols are present in the sequence of symbols, wherein a signal is a window of the sequence of symbols consisting of a predefined number of contiguous symbols;

wherein the sequence of amino acids is analyzed from the identity of the signals present in the sequence of symbols.

2. The method of claim 1, wherein the window consists of 5-15 contiguous symbols.

3. The method of claim 1, wherein the window consists of 9 contiguous symbols.

4. The method of claim 1, wherein the predetermined set of amino acids consists of 4-10 amino acids, and at least 4 are selected from the group consisting of A, R, Q, E, L, K and M.

5. The method of claim 4, wherein the predetermined set of amino acids consists of A, R, Q, E, L, K and M.

6. The method of claim 1, wherein the predetermined set of amino acids consists of 4-10 amino acids, and at least 4 are selected from the group consisting of C, I, L, M, F, W, Y, and V.

7. The method of claim 6, wherein the predetermined set of amino acids consists of C, I, L, M, F, W, Y, and V.

8. The method of claim 1, further comprising transforming the sequence of symbols into a sequence of signal designations, wherein different designations are used to represent different signals in the sequence of symbols.

9. The method of claim 1, wherein an amino acid is designated with a first type of second symbol if it is part of a second predetermined set of amino acids, and a second type of second symbol if it is not part of the second set of amino acids.

10. The method of claim 1, wherein the signals present in the sequence of symbols are assigned grades according to the probability that the observed frequency of a signal in a collection of proteins in which each amino acid has been designated with a symbol occurs by chance, wherein the grade increases with decreasing probability.

11. The method of claim 10, wherein the signals are classified as significant or not significant signals depending whether the grade exceeds a threshold.

12. The method of claim 11, wherein the threshold is a $\chi^2 > 8$ that the observed frequency of the signal in the collection of proteins does not occur by chance.

13. The method of claim 12, further comprising determining the number and identity of significant signals in the amino acid sequence.

14. The method of claim 13, wherein the sequence of amino acids is a theoretical amino acid sequence, and the method further comprises determining the probability that the theoretical amino acid sequence is an actual protein by comparing the expected number of significant signals in the theoretical amino acid sequence to the actual number of significant signals in the theoretical amino acid sequence.

15. The method of claim 14, wherein the theoretical amino acid sequence is designated as an actual protein sequence if the probability that the observed significant signals in the sequence arose by chance is 10^{-10} or less.

16. The method of claim 1, wherein the sequence of amino acids is from a known protein.

17. The method of claim 1, wherein the sequence of amino acids is from a putative protein.

18. The method of claim 1, further comprising repeating steps (a) and (b) for a second sequence of amino acids and aligning the sequences of symbols produced from the first and second sequences of amino acids for maximum conservation of significant signals.

19. The method of claim 11, further comprising predicting the secondary structure of a segment of a protein located within the sequence of amino acids from the identity of significant signals.

20. The method of claim 19, wherein the secondary structure is selected from the group consisting of an alpha helix, beta strand, beta turn, turn + beta, helix + turn, helix cap, extended helix, Gly/Pro twist, beta + turn, helix-hairpin, beta cap, helix hairpin, beta hairpin, contorted helix, turn, helix + turn II and helix turn.

21. The method of claim 1, further comprising inputting the sequence of amino acids into the computer.

22. The method of claim 21, wherein the sequence of amino acids is input by transfer of data from a database.

23. The method of claim 1, further comprising outputting the identity of signals present in the sequence of symbols.

24. The method of claim 23, wherein the signals are output in an order corresponding to the order of amino acids in the sequence of amino acids.

25. The method of claim 1, further comprising providing user input of the predefined number of contiguous symbols of the window.

26. The method of claim 10, further comprising calculating the probability that the observed frequency of a signal in the collection of proteins in which each amino acid has been designated with a symbol occurs by chance.

27. The method of claim 1, wherein step (b) determines the identity of L- (P-1) signals within the sequence of amino acids, where L is length of the sequence of amino acids and P is the predefined number of contiguous symbols in the window.

28. The method of claim 1, further comprising assigning the determined signals designations, a different designation being used for each unique signal.

29. The method of claim 1, further comprising analyzing the sequence of amino acids from the identity of the signals.

30. A computer implemented method of identifying a set of amino acids useful for the analysis of proteins, comprising

(a) designating each amino acid within each of a collection of proteins with a symbol, wherein an amino acid is designated a first symbol if it is a member of a first test set of amino acids, and a second symbol different from the first symbol if the amino acid is not a member of the test set, thereby producing a collection of sequences of symbols;

(b) determining the number of occurrences of different signals of the symbols in the collection of sequences of symbols, wherein a signal is a window of the sequence of symbols consisting of a predefined number of contiguous symbols; and

(c) determining the probability that the distribution of the number of signals of each signal strength occurs by chance, wherein the lower the probability the more useful the test set of amino acids is for protein analysis.

31. The method of claim 30, further comprising repeating steps (a), (b) and (c) for a second test set of amino acids.

32. The method of claim 31, wherein the second test set differs from the first test set by the addition, deletion, or substitution of an amino acid from the first test set.

33. The method of claim 32, further comprising repeating steps (a), (b) and (c) for each possible unique set of amino acids consisting of 4-10 amino acids.

34. The method of claim 1, further comprising comparing the position and identity of each signal present in the sequence of symbols to a conserved signal pattern present in a family of proteins.

35. A computer-implemented method of predicting the fold of a query protein comprising;

(a) designating each amino acid within a family of protein sequences with a symbol, wherein an amino acid is designated a first symbol if it is a member of a predetermined set of amino acids, and a second symbol different from the first symbol if the amino acid is not a member of the set, thereby producing a plurality of sequences of symbols;

(b) determining which signals of the symbols are present in the sequences of symbols, wherein a signal is a window of the sequence of symbols consisting of a predefined number of contiguous symbols;

(c) determining a conserved signal pattern between members of the family;

(d) analyzing a query protein to identify a signal pattern;

(e) determining if the query protein's signal pattern exceeds a threshold of similarity to the conserved signal pattern; and

(f) if the signal pattern of the query exceeds the threshold, designating the query as having the fold of the family.

36. The method of claim 35, further comprising comparing the query protein's signal pattern to conserved signal patterns in an additional protein family.

37. The method of claim 36, wherein the family is selected from the list consisting of globins, lysozymes, thioredoxins, trypsins, monoclonal antibodies, and amido transferases.

38. The method of claim 35, wherein the conserved signal pattern includes a signal present in Table 14.

39. The method of claim 35, wherein the conserved signal pattern includes a signal present in Table 15.

40. The method of claim 5, wherein at least one signal present in the sequence of symbols is present in Table 14.

41. The method of claim 7, wherein at least one signal present in the sequence of symbols is present in Table 14.

42. The method of claim 5, wherein at least one signal present in the sequence of symbols is present in Table 15.

43. The method of claim 7, wherein at least one signal present in the sequence of symbols is present in Table 15.

44. The method of claim 14, wherein at least one signal present in the sequence of symbols is present in Table 14.

45. The method of claim 14, wherein at least one signal present in the sequence of symbols is present in Table 15.

46. A computer program product stored on a computer readable media for analyzing a sequence of amino acids, the program product comprising;

(a) code for designating each amino acid within the sequence with a symbol, wherein an amino acid is designated a first symbol if it is a member of a predetermined set of amino acids, and a second symbol different from the first symbol if the amino acid is not a member of the set, thereby producing a sequence of symbols; and

(b) code for determining which signals of the symbols are present in the sequence of symbols, wherein a signal is a window of the sequence of symbols consisting of a predefined number of contiguous symbols, wherein the sequence of amino acids is analyzed from the identity of the signals present in the sequence of symbols.

47. A computer program product stored on a computer readable media for identifying a set of amino acids useful for the analysis of proteins, the program product comprising:

(a) code for designating each amino acid within each of a collection of proteins with a symbol, wherein an amino acid is designated a first symbol if it is a member of a first test set of amino acids, and a second symbol different from the first symbol if the amino acid is not a member of the test set, thereby producing a collection of sequences of symbols;

(b) code for determining the number of occurrences of different signals of the symbols in the collection of sequences of symbols, wherein a signal is a window of the sequence of symbols consisting of a predefined number of contiguous symbols; and

(c) code for determining the probability that the distribution of the number of signals of each signal strength occurs by chance, wherein the lower the probability the more useful the test set of amino acids is for protein analysis.

48. A computer program product stored on a computer readable media for predicting the fold of a query protein, the program product comprising:

(a) code for designating each amino acid within a family of protein sequences with a symbol, wherein an amino acid is designated a first symbol if it is a member of a predetermined set of amino acids, and a second symbol different from the first symbol if the amino acid is not a member of the set, thereby producing a plurality of sequences of symbols;

(b) code for determining which signals of the symbols are present in the sequences of symbols, wherein a signal is a window of the sequence of symbols consisting of a predefined number of contiguous symbols;

(c) code for determining a conserved signal pattern between members of the family;

(d) code for analyzing a query protein to identify a signal pattern;

(e) code for determining if the query protein's signal pattern exceeds a threshold of similarity to the conserved signal pattern; and

(f) code for designating the query as having the fold of the family if the signal pattern of the query exceeds the threshold.

49. A computer program product stored on a computer readable media for identifying a coding region of a nucleotide sequence, the program product comprising:

(a) code for translating all possible reading frames of a nucleotide sequence into theoretical protein sequences;

(b) code for designating each amino acid within the theoretical protein sequences with a symbol, wherein an amino acid is designated a first symbol if it is a member of a first predetermined set of amino acids, and a second symbol different from the first symbol if the amino acid is not a member of the predetermined set, thereby producing a collection of sequences of symbols;

(c) code for determining the number of significant signals in each reading frame of the nucleotide sequence; and

(d) code for determining an expected number of significant signals in each reading frame of the nucleotide sequence.

50. A system for analyzing a sequence of amino acids, comprising:

(a) a processor; and

(b) a memory coupled to the processor configured to store a plurality of instructions which when executed by the processor cause the processor to:

(i) designate each amino acid within the sequence with a symbol, wherein an amino acid is designated a first symbol if it is a member of a predetermined set of amino acids, and a second symbol different from the first symbol if the amino acid is not a member of the set, thereby producing a sequence of symbols;

(ii) determine which signals of the symbols are present in the sequence of symbols, wherein a signal is a window of the sequence of symbols consisting of a predefined number of contiguous symbols, wherein the sequence of amino acids is analyzed from the identity of the signals present in the sequence of symbols.

51. A system for identifying a set of amino acids useful for the analysis of proteins comprising:

(a) a processor; and

(b) a memory coupled to the processor configured to store a plurality of instructions which when executed by the processor cause the processor to:

(i) designate each amino acid within each of a collection of proteins with a symbol, wherein an amino acid is designated a first symbol if it is a member of a first test set of amino acids, and a second symbol different from the first symbol if the amino acid is not a member of the test set, thereby producing a collection of sequences of symbols;

(ii) determine the number of occurrences of different signals of the symbols in the collection of sequences of symbols, wherein a signal is a window of the sequence of symbols consisting of a predefined number of contiguous symbols; and

(iii) determine the probability that the distribution of the number of signals of each signal strength occurs by chance, wherein the lower the probability the more useful the test set of amino acids is for protein analysis.

52. A system for predicting the fold of a query protein comprising:

(a) a memory;

(b) a system bus;

(c) a processor operatively disposed to:

(i) designate each amino acid within a family of protein sequences with a symbol, wherein an amino acid is designated a first symbol if it is a member of a predetermined set of amino acids, and a second symbol different from the first symbol if the amino acid is not a member of the set, thereby producing a plurality of sequences of symbols;

(ii) determine which signals of the symbols are present in the sequences of symbols, wherein a signal is a window of the sequence of symbols consisting of a predefined number of contiguous symbols;

(iii) determine a conserved signal pattern between members of the family;

(iv) analyze a query protein to identify a signal pattern;

(v) determine if the query protein's signal pattern exceeds a threshold of similarity to the conserved signal pattern; and

(vi) designate the query as having the fold of the family if the signal pattern of the query exceeds the threshold.

53. A system for identifying a coding region of a nucleotide sequence comprising:

(a) a memory;

(b) a system bus;

(c) a processor operatively disposed to:

(i) translate all possible reading frames of a nucleotide sequence into theoretical protein sequences;

(ii) designate each amino acid within the theoretical protein sequences with a symbol, wherein an amino acid is designated a first symbol if it is a member of a first predetermined set of amino acids, and a second symbol different from the first symbol if the amino acid is not a member of the predetermined set, thereby producing a collection of sequences of symbols;

(iii) determine the number of significant signals in each reading frame of the nucleotide sequence; and

(iv) determine an expected number of significant signals in each reading frame of the nucleotide sequence.